CLAIMS

What is claimed is:

SUB B33

1. A method for uniform representation of a subject genome sequence comprising the steps of:

providing a set of known biological fragments, the set being of a predetermined number of said known biological fragments;

comparing each known biological fragment from the set to a subject genome sequence, for each known biological fragment said comparing including (i) counting the number of times the known biological fragment is found in the subject genome sequence and (ii) from said counted number of times, forming a vector element, such that for each known biological fragment there is a respective vector element representing the number of times that known biological fragment is found in the subject genome sequence; and

from the formed vector elements, forming a vector having a length equal to the predetermined number of known biological fragments in the provided set, such that the formed vector provides a fixed length representation of the subject genome sequence.

2. A method as claimed in Claim 1 wherein the set of known biological fragments is from published databases of motifs or proteins.

3. A method as claimed in Claim 1 further comprising the step of:

SUB B4

for each desired subject genome sequence, repeating the comparing and forming steps such that a respective vector representation is formed and each desired subject genome sequence has a same length vector representation.

4. A method as claimed in Claim 3 wherein for each subject genome sequence, having formed respective vector representations each of the same length, using the same length vector representation as input into one or more sequence analyses.

5. A method as claimed in Claim 4 wherein the sequence analyses include one of indexing, classification and clustering.

6. A method as claimed in Claim 1 wherein the subject genome sequence is a protein sequence or subsequence.

7. A method as claimed in Claim 1 wherein the subject genome sequence is a DNA sequence or subsequence.

8. A method as claimed in Claim 1 wherein the counting includes determining probability of the subject genome sequence being generated by the known biological fragment.

9. A method as claimed in Claim 8 wherein the counting determining probability employs a 0th order Markov model for each known biological fragment.

10. Apparatus for forming uniform representations of genome sequences, comprising:

a data store of a predefined number of known biological sequences;

a comparison routine executed by a digital processor having access to the data store, the comparison routine comparing each known biological sequence from the data store to a subject genome sequence and generating a score indicative of the comparison, said scores forming a vector having a length equal to the predefined number of known biological sequences, such that said

comparison routine outputs the formed vector as a fixed length representation of the subject genome sequence.

11. Apparatus as claimed in Claim 10 wherein the data store is a published database of motifs or proteins.

12. Apparatus as claimed in Claim 10 further comprising a plurality of different subject genome sequences; and

wherein the comparison routine forms for each subject genome sequence, a respective vector such that a corresponding plurality of same length vector representations is provided.

13. Apparatus as claimed in Claim 12 wherein the output of the comparison routine feeds the corresponding plurality of same length vector representations into further analysis processors.

14. Apparatus as claimed in Claim 13 wherein the further analysis processors include at least one of a classifier, an indexer and a clustering member.

15. Apparatus as claimed in Claim 10 wherein the subject genome sequence is a protein sequence or subsequence.

16. Apparatus as claimed in Claim 10 wherein the subject genome sequence is a DNA sequence or subsequence.

17. Apparatus as claimed in Claim 10 wherein the generated score is a probability of the subject genome sequence being generated by the known biological sequence.

18. Apparatus as claimed in Claim 10 wherein the generated score is a counting of a number of occurrences of the known biological sequence found in the subject genome sequence.

ADD B2